Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies

Fortier, Isabel; Doiron, Dany; Little, Julian; Ferretti, Vincent; L'Heureux, Francois; Stolk, Ronald P.; Knoppers, Bartha M.; Hudson, Thomas J.; Burton, Paul R.; Int Harmonization Initiative

Published in:
International Journal of Epidemiology

DOI:
10.1093/ije/dyr106

Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2011;**40**:1314–1328
doi:10.1093/ije/dyr106

Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies

Isabel Fortier,[1,2]* Dany Doiron,[2] Julian Little,[1,3] Vincent Ferretti,[4] François L'Heureux,[2] Ronald P Stolk,[5] Bartha M Knoppers,[1,6] Thomas J Hudson,[4,7,8] and Paul R Burton[2,9,10] on behalf of the International Harmonization Initiative[†]

[1]Research Institute – McGill University Health Centre, Montreal, Quebec, Canada, [2]Public Population Project in Genomics (P^3G), Montreal, QC, Canada, [3]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, ON, Canada, [4]Ontario Institute for Cancer Research, MaRS Centre, Toronto, ON, Canada, [5]Department of Epidemiology, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands, [6]Department of Human Genetics, Centre of Genomics and Policy, Faculty of Medicine, McGill University, Montreal, QC, Canada, [7]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada, [8]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada and [9]Department of Health Sciences, University of Leicester, Leicester, UK and [10]Department of Genetics, University of Leicester, Leicester, UK

*Corresponding author. Research Institute – McGill University Health Centre, Allen Memorial Building, 1025 Pine Avenue West, room P2.028, Montreal, Quebec, Canada, H3A 1A1. E-mail: ifortier@p3g.org
[†]The members of the International Harmonization Initiative are provided in Appendix.

Accepted 31 May 2011

Background Proper understanding of the roles of, and interactions between genetic, lifestyle, environmental and psycho-social factors in determining the risk of development and/or progression of chronic diseases requires access to very large high-quality databases. Because of the financial, technical and time burdens related to developing and maintaining very large studies, the scientific community is increasingly synthesizing data from multiple studies to construct large databases. However, the data items collected by individual studies must be inferentially equivalent to be meaningfully synthesized. The DataSchema and Harmonization Platform for Epidemiological Research (DataSHaPER; http://www.datashaper.org) was developed to enable the rigorous assessment of the inferential equivalence, i.e. the potential for harmonization, of selected information from individual studies.

Methods This article examines the value of using the DataSHaPER for retrospective harmonization of established studies. Using the DataSHaPER approach, the potential to generate 148 harmonized variables from the questionnaires and physical measures collected in 53 large population-based studies (6.9 million participants) was assessed. Variable and study characteristics that might influence the potential for data synthesis were also explored.

Results Out of all assessment items evaluated (148 variables for each of the 53 studies), 38% could be harmonized. Certain characteristics of variables (i.e. relative importance, individual targeted, reference period) and of studies (i.e. observational units, data collection start date and mode of questionnaire administration) were

associated with the potential for harmonization. For example, for variables deemed to be essential, 62% of assessment items paired could be harmonized.

Conclusion The current article shows that the DataSHaPER provides an effective and flexible approach for the retrospective harmonization of information across studies. To implement data synthesis, some additional scientific, ethico-legal and technical considerations must be addressed. The success of the DataSHaPER as a harmonization approach will depend on its continuing development and on the rigour and extent of its use. The DataSHaPER has the potential to take us closer to a truly collaborative epidemiology and offers the promise of enhanced research potential generated through synthesized databases.

Keywords Data synthesis, data quality, data pooling, harmonization, meta-analysis, DataSHaPER, retrospective harmonization

Introduction

In order to properly understand the role of, and interaction among genetic, lifestyle, environmental and social factors in modulating the risk of development and/or progression of chronic diseases, it is critical that analyses have adequate statistical power.[1,2] Moreover, the aetiological relations of interest are generally complex and the risks targeted often relatively weak.[1,3] Scientific progress thus demands access to very large databases[4] providing comprehensive, valid and precise information on a variety of factors and disease traits.[5] Due to the financial, technical and time burdens related to developing and maintaining very large studies, the scientific community is increasingly making use of the synthesis of data to address the limitations of statistical power of individual studies.[6] In turn, however, when an analysis is to be undertaken based on individual-level data integrated between studies, the key data items assessed in individual studies (e.g. specific measurements or questions) must not only be measured well, but the information they convey must also be rendered 'inferentially equivalent' (or 'harmonized').[7] A fundamental tension therefore lies between increasing sample sizes by synthesizing data from many individual studies, and restriction of the data synthesis to those studies that are satisfactorily harmonized and which provide a common set of scientifically valid information.

Given the quantity and complexity of the information generally collected by individual established studies and the heterogeneity of their designs and procedures, it is essential to have access to effective methods and tools to formally explore the potential to generate high-quality synthesized databases. Such tools need to provide comprehensive information on: (i) the specific variables that could be shared; (ii) the studies that could participate in targeted analysis; and (iii) the factors that could influence the potential to integrate information. The DataSchema and Harmonization Platform for Epidemiological Research (DataSHaPER; www.datashaper.org) may be used to provides such information.[7] It includes two primary components. The DataSchema[8] documents and annotates sets of core variables, which each provides a concise but effective list of information to be harmonized in a specific scientific context. The Harmonization Platform then provides a template for the formal estimation of the potential to synthesize information across networks of studies. In collaboration with other Public Population Project in Genomics (P^3G) partner teams[9–12] and projects,[13,14] developments are currently ongoing to build a suite of tools supporting all steps of individual data harmonization, synthesis and analysis.

The current article evaluates the value of the DataSHaPER as a foundation for retrospective harmonization. It focuses on the Generic DataSchema that is composed of variables aimed at supporting the construction of general purpose baseline questionnaires and physical measures for use in large population-based studies enrolling adult participants.[7] The specific aim of the article is to investigate and quantify the potential for harmonizing variables that make up the Generic DataSchema questionnaires and physical measures modules[7] across more than 50 of the world's largest population-based studies. The article also investigates factors that might influence the potential for harmonization, particularly the basic study design and nature of individual variables, as well as evaluating the utility of the Generic DataSchema as a platform for harmonization of existing databases. The article will help guide future development of the DataSHaPER project.

Methods

Selection of variables

Development of the Generic DataSchema has been centred on a series of international consensus workshops bringing together experts from more than 25 studies and 14 countries. Its construction was based upon iterative review and consensus methodologies[15,16] aimed at synthesizing knowledge and input from practitioners and researchers with a variety of professional expertise. The primary objective was to generate a select list of core variables that would provide a valuable contribution to the baseline interviews of population-based studies enrolling adult participants and provide a basis for harmonization. Agreed selection criteria were defined and used, first, to select broad domains and, then, specific variables.[7] Where possible, variables were chosen and defined so as to be compatible with widely used international classification systems or standards.[17–19] Over a 3-year period, iterative rounds of discussion and comment contributed to gradual refinement of the Generic DataSchema. In particular, a number of cohorts provided critical feedback based on their own use of early versions of the DataSchema as a foundation for questionnaire development. These included Lifelines[20] (The Netherlands), LifeGene[21] (Sweden), the five cohorts in the Canadian Partnership for Tomorrow Project[14] and the Canadian Longitudinal Study on Aging.[22] The Generic DataSchema remains a dynamically evolving tool that is improving incrementally over time based on users' feedback and increasing scientific knowledge.[7]

The ontology of the Generic DataSchema has a hierarchical structure in which the selected variables (the fundamental units of a statistical analysis, such as: lifetime occurrence of cancer; highest level of education; diastolic blood pressure) are grouped under domains (e.g. individual history of cancer, education level, blood pressure) that are themselves distributed across themes (e.g. individual history of disease, socio-economic status, body function measures) and modules (health and risk factor questionnaire, physical and cognitive measures and interview administration).[7] Only the health and risk factor questionnaire and the physical and cognitive measures modules were considered for harmonization. Basic interview administration information (e.g. date of birth, date of interview, sex, language and location of interview) was collected by participating studies but not necessarily through collection modes targeted by the current exercise. Hence, even though all studies could share constructs derived from these data (such as age), this module was excluded from the current harmonization exercise. Within the physical and cognitive measures module, contraindications to having a particular measurement were also excluded from the harmonization process, since we did not have access to consistent information on the specific contraindications targeted by studies.

Selected variable characteristics that might influence the potential for data synthesis were documented. These include: (i) its importance for broad-based research projects as perceived by expert consensus (essential, important, useful information), (ii) the individual targeted by the variable (the participant himself/herself, the participant's family members) and (iii) the period of the participant's life to which the variable refers (current status at assessment, all other periods). Classification of the perceived 'importance of variables' was based on their relevance to be part of a generic set of information useful for a broad range of research questions. Since this classification was subjective, it was undertaken by two separate experts followed by comparison, and final confirmation by a panel. As an illustrative example, the following variables were classified as: 'essential': occurrence of cancer (participant); 'important': occurrence of cancer in the family; 'useful': occurrence of cancer in siblings.

Selection of participating studies

Potential studies to be included in the harmonization process were identified using P^{3}G's collaborative network and Study catalogue (www.p3gobservatory.org). Eligible studies were those that had: (i) recruited or planned to recruit at least 10 000 adult participants; (ii) collected biological samples enabling DNA extraction; (iii) collected comprehensive information on life habits, socio-economic status and health outcomes; and (iv) provided access to the baseline questionnaire and standard operating procedures used. Priority was given to studies with the largest number of participants, those that were initiated most recently and those that had questionnaires available in English or French. A total of 87 studies were initially approached. Of these, 53 agreed to participate; 6 declined and 28 did not respond to our contacts.

Key characteristics were documented in order to identify study-related factors that might influence the potential for synthesis. These were: (i) study design (cohort or cross-sectional); (ii) breadth of scientific focus (broad scientific focus or focus on specific outcomes or exposures); (iii) source of population sample (general population or specific subpopulation such as clinic out-patients, professional association, etc.); (iv) participation history (new participant or participant selected from a pre-existing study); (v) observational units (individuals or families); (vi) sex (whether one sex only or both sexes were included); (vii) region of residence of the participants (Europe, North America or other); (viii) lower limit of participants' age at recruitment (<25, 25–40 or >40 years); (ix) data collection start date (before or in 2000 or after 2000), (x) targeted number of participants (less than 100 000 or 100 000 or more); (xi) proportion of participants from whom biological samples were collected (a subsample or all participants); (xii) mode of administration of the

Box 1 Classification of the level of compatibility between assessment items and DataSchema variables. (For a detailed example, please refer to Fortier *et al.*[7])

Class[a]	Description
Complete	According to the pairing rules, the meaning, format and standard operating procedures used for collection of the assessment items allow construction of the variable as defined.
Partial	According to the pairing rules, the meaning, format and standard operating procedures used for collection of the assessment items allow the construction of the variable as defined, but with an unavoidable loss of information. This class includes two subcategories:
	'Proximate'—when the only reason for the classification as partial is because categories are used to collect information for a DataSchema variable that is defined as continuous.
	'Tentative'—whenever a variable is classified as partial for any other reason.
Impossible	When no relevant information is collected (impossible not covered) or, based on the pairing rules, insufficient information exists to construct the variable as defined (impossible covered).

[a]In certain instances, a DataSchema variable is not pertinent in the context of a particular study (e.g. the 'Occurrence prostate cancer' variable in the context of a study recruiting only women). In such cases, the variable is classified as 'Not applicable' for that study.

questionnaire (computer based, paper based or mixed); and (xiii) site of administration of the questionnaire (participant residence, clinic/assessment centre or mixed).

Harmonization process

Using the 53 participating studies, an initial evaluation was undertaken in order to determine if the selected Generic DataSchema domains correspond to risk factors and outcomes most frequently collected in current practice. This was done using the keyword tree of the P^3G questionnaire catalogue, which provides a comprehensive overview of the areas of information collected by large epidemiologic studies.[23] Each DataSchema domain was attributed a corresponding keyword of the keyword tree [e.g. 'DataSchema domain', Individual history of Cancer; 'Keyword', Individual History of Neoplasm—International Classification of Diseases (ICD)]. The proportion of the 53 participating studies collecting information on each keyword (i.e. coverage) was determined. This then allowed examination of the extent to which areas of information that were selected to be part of the Generic DataSchema (i.e. keywords corresponding to DataSchema domains), or not selected (i.e. keywords not included in the DataSchema), were used in current practice. The result of this analysis will help to optimize the Generic DataSchema by adding or deleting specific domains of interest based on their coverage among the world's leading population-based studies.

As the primary harmonization exercise, the potential for each study to generate each variable included in the Generic DataSchema was evaluated. Once specific definitions have been attributed to each variable, the harmonization approach entails using a three-level scale to document the compatibility between the information generated by the assessment items of each participating study (e.g. specific questions) and each variable defined in the DataSchema. This process is referred to as 'pairing' (Box 1).[7] In order to classify the assessment items and to ensure the validity and reproducibility of the pairing results, sets of comprehensive 'pairing rules' specific to each variable are defined.[7] Development of pairing rules is context specific and involves a systematic process of iteration between scientific experts and trained research officers. Using these pairing rules, trained research officers determine whether or not a variable can be recreated using the assessment items collected by each participating study. In order to ensure quality control, an evaluator verified a random sample of ~10% of all pairing results. Whenever a pairing error was detected, a consensus panel determined the final pairing result.

Statistical methods

The chi-square statistic was used to test for difference between the characteristics of participating and non-participating studies. Study and variable characteristics were documented using descriptive statistics for binary and nominal variables including tabulation and reporting of counts and percentages. Coverage of areas of information that had been selected or not selected to be part of the Generic DataSchema were compared using the non-parametric Wilcoxon rank sum test.[24] Cohen's κ statistic[25] was used to evaluate agreement between the research officers conducting the initial pairing and the final validated results. Primary analysis was based on the standard unweighted κ statistic[25] but, because the classification categories were ordered, two alternatively weighted κ statistics[26] were also used.

Analysis of the study- and variable-level characteristics that systematically influence the likelihood that a given variable can be created from the assessment

items of a given study was undertaken. This analysis was used to identify factors that influence potential for data synthesis and was based upon a type 1 generalized estimating equations (GEE) analysis,[27,28] which takes appropriate account of the correlation of variables within studies. The analysis was based on a logistic regression model with an exchangeable correlation structure. Robust standard errors were used throughout.

Results

Participating studies

Among the 87 studies contacted, no association was observed between study size or region of origin and the likelihood of participation in the initiative. However, studies that launched recruitment of participants after 2000 were more likely to respond to our call to participate [$P < 0.05$, odds ratio (OR) = 2.73, 95% confidence interval (CI) = 1.11–6.71]. Fifty-three studies accepted to participate, encompassing 6.9 million participants, recruited or to be recruited. Of the participating studies (Supplementary Data available at *IJE* online), 48 are designated as cohorts and 5 had cross-sectional designs. Most studies had a broad scientific focus (42 studies), enrolled new participants (46 studies), had the individual—rather than family—as the unit of observation (47 studies) and included both sexes (46 studies). Participants were recruited from the general population in most studies (44 studies). However, some (nine studies) targeted population subgroups such as professional associations, clinic out-patients and at-risk individuals (exposure or disease). The start dates for collection of information range from the mid-1980s to 2010. Five studies began collecting information in 1990 or before and 18, 10 and 20, respectively, in 1991–2000, 2001–05 and 2006–10. Study sizes range from 10 000 to over half a million participants. Eight have already recruited, or aim to recruit, 500 000 participants or more, whereas the majority (28 studies) have a participant recruitment target of 50 000 or less. Thirty studies have collected or plan to collect biological samples from all participants, whereas the others collect samples from a subgroup of participants only. A total of 26 studies recruit participants from Europe, 17 from North America, 5 from Asia and 5 from Australia. Different data collection modes are also used. For example, 24 studies are using paper-based questionnaires and 13 used computer-based questionnaires; the rest used a mixed data collection format.

Generic DataSchema

Areas of information

In order to confirm that the Generic DataSchema does provide effective coverage of areas of interest to a general purpose population-based study, and in order to assess its use in actual current practice, all 53 participating studies were keyword coded. The keywords themselves were categorized into those that do correspond to a Generic DataSchema domain and those that do not. The proportion of studies that covered each keyword was then calculated and plotted (Figure 1). Keywords that do correspond to DataSchema domains are covered, on average, by 75.9% of studies, whereas keywords that do not are covered by 25%. The mean difference of coverage was therefore estimated at 50.9% ($P < 0.00001$, 95% CI = 42.2–59.2). Furthermore, all keywords covered by at least 78% of participating studies correspond to DataSchema domains (Figure 1). However, even if the DataSchema includes areas generally used in current practice, a few exceptions are noted. For example, the 'Subject family's birth location' keyword is included in the DataSchema, but is only covered by 34% of the studies. Similarly, 'Familial diseases of the respiratory system' and 'Familial mental and behavioural disorders' (ICD10) keywords are also in the DataSchema, but only covered by 36 and 38% of the studies, respectively. On the other hand, 'Individual disease of the genitourinary system' (ICD10) and 'Individual disease of the digestive system' (ICD10) are covered, respectively, by 75 and 77% of the studies, but are not included in the DataSchema.

Variables

All 148 Generic DataSchema variables encompassed by the questionnaire and physical measures modules were included in the pairing exercise. Among the variables paired, 38 were classified as 'essential' for broad focused epidemiological research, 45 were considered 'important' and 65 were deemed 'useful'. A total of 101 variables (68%) target information directly related to the participant (e.g. Occurrence of menopause, Height), whereas 47 (32%) target the participant's family (e.g. Occurrence of diabetes in siblings). Forty variables (27%) target information related to the time at which the assessment took place (e.g. Current smoking status), whereas the remaining 108 variables target information relating to other time periods in the participant's life (e.g. Age at first pregnancy).

Harmonization results

The overall pairing process shows that 36% of all assessment items paired (148 variables paired for each of the 53 studies) were classified as 'Complete', 3% as 'Partial Proximate', 14% as 'Partial Tentative', 47% as 'Impossible' (7.2% 'Impossible Covered' and 40.2% 'Impossible Not Covered) and 0.2% as 'Not Applicable'. See Box 1 for definitions of pairing classifications. Some continuous variables, defined as open-ended in the Generic DataSchema, simultaneously show a very low proportion of 'Complete' matches and a relatively high percentage of 'Partial Proximate' matches (e.g. Household income—4%

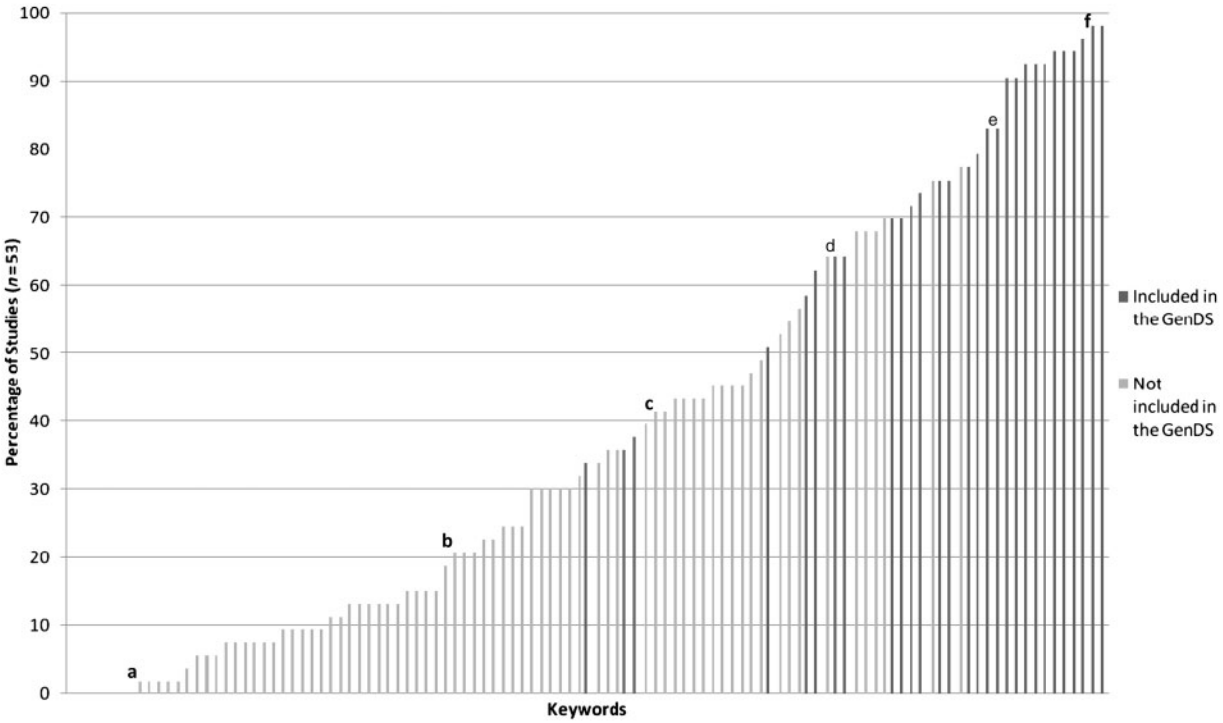

Figure 1 Percentage of studies covering keywords included and not included in the Generic DataSchema. Example of keywords not included in the Generic DataSchema (percentage of studies covering): (a) Certain conditions originating in the perinatal period (2%); (b) Language (21%) and (c) Certain infectious and parasitic diseases (42%). Example of keywords included in the Generic DataSchema (percentage of studies covering): (d) Endocrine, nutritional and metabolic diseases (64%); (e) Diseases of the respiratory system (83%) and (f) Tobacco use (98%)

Table 1 Pairing results (%) for selected variables presenting a high proportion of 'Complete/Partial Proximate' matches*

Variable name	Complete/Partial proximate (%)
Occurrence of diabetes	89
Current use of alcohol	89
Level of physical activity	85
Occurrence of high blood pressure	83
Occurrence of menopause	83
Occurrence of cancer	81
Occurrence of stroke	81
Employment status	81
Menopause onset	75
Occurrence of asthma	74
Current quantity of cigarettes smoked	74
Occurrence of myocardial infarction	72
Living with partner	68
Type of cancer	66
Standing height	66
Weight	66

*For full-pairing results, please refer to supplementary materials (Supplementary Data available at *IJE* online).

'Complete' matches and 47% 'Partial Proximate' matches). Although such variables stray from the ideal 'Complete' match because of the use of categorical responses in study questionnaires, they can nonetheless be synthesized across studies with a minor loss of information. In the following results, 'Partial Proximate' matches are grouped with 'Complete' matches. Across the 53 studies, the proportion of combined 'Complete' and 'Partial Proximate' pairings ranged from a minimum of 20% (29 out of 148 variables) up to a maximum of 90% (133 out of 148 variables) per individual study. Conversely, the variation across the DataSchema variables ranged between 4% (2 of 53 studies) and 89% (47 of 53 studies). For 35 variables, at least 60% of the participating studies provided a 'Complete/Partial Proximate' match (Table 1 for examples of 'well-paired' variables, i.e. 65% or more). On the other hand, for 12 of the 148 variables (e.g. Heaviest ever alcohol consumption, Occurrence of chronic obstructive pulmonary disease in the family, Town of birth), at least 80% of the participating studies provided an 'Impossible not covered' match, meaning these variables were in fact not examined by the majority of participating studies. Complete pairing results of all Generic DataSchema variables are included in Supplementary materials (Supplementary Data available at *IJE* online).

Although we have deliberately chosen to present the data in a conservative manner (so that we are not overstating the quality of pairing), it might reasonably be argued that evaluation of the harmonization potential should sometimes be restricted to those studies that have actually collected information related to a particular variable of interest. That being the case, each 'Impossible Not Covered' and 'Not Applicable' matches would be excluded from the denominator. This increases the overall proportion of 'Complete/Partial Proximate' matches to 64% and limits the 'Impossible' matches to those that had insufficient information to construct the variable ('Impossible Covered'; 12%).

For quality control purposes, a total of 718 randomly paired assessment items were subjected to blind reassessment by a validation panel (9% of all assessment items paired). A total of 3.2% of the validated pairing matches indicated a disagreement between the original assessor and the validation panel in categorizing the pairing as 'Complete', 'Partial' or 'Impossible'. Cohen's κ was estimated and indicated excellent agreement between the original pairing assessments and the validation panel's reassessments (OR = 0.948, 95% CI = 0.926–0.970). The occurence of pairing errors was not influenced by the nature of the variable, the research officer conducting pairing, or other characteristics such as paring levels. Since the allocation categories were ordered, a sensitivity analysis using two alternatively weighted κ statistics was also undertaken—the results were very similar: OR: 0.945 (95% CI: 0.923–0.967) and OR: 0.957 (95% CI: 0.939–0.975).

Study-specific pairing results

Table 2 presents the association between pairing results and study characteristics. The study's observational units, data collection start date, proportion of biological samples collected and mode of questionnaire administration are each associated with pairing results in a series of univariate models. The table also presents the variables that were associated with pairing results using a multivariate model (observational units, data collection start date and mode of questionnaire administration). When considering only the nine studies that exhibit characteristics that typically facilitate harmonization (i.e. those enrolling individual participants, not families; commencing after the year 2000 and using computerized questionnaires), the overall proportion of 'Complete/Partial Proximate' matches was 50%. This proportion is higher than the 38% seen when all 53 studies are considered.

Variable-specific pairing results

Three variable-specific characteristics exhibited substantive association with the potential for harmonization (Table 3): variable importance; individual targeted by the variable; period targeted by the variable. About 62% of pairing matches for variables considered as collecting 'essential' information were 'Complete/Partial Proximate', compared with 43% for variables considered 'important' and 22% for variables classified as 'useful'. For the nine studies outlined in the section above, when only the 'essential' variables are considered (38 variables), 81% of pairing matches are either 'Complete' or 'Partial Proximate'.

Case study

A case study is used to illustrate the potential scientific utility of harmonization under the DataSHaPER. It was defined during a workshop involving participating studies and explores the possibility to investigate the association between blood pressure and five risk factors: education level; body mass index; physical activity; current alcohol use and cigarette consumption (Table 4). Variables were initially chosen based on their scientific relevance, but harmonization potential was also considered in order to finalize selection. As noted earlier, although age and sex variables were not included in the formal harmonization process, they were nonetheless available for all participating studies and could be included as covariates in this pooled analysis.

Based on the pairing results, 14 studies (~2 million participants) could potentially collaborate fully in a pooled analysis of associations between blood pressure and the stated factors. For the 14 studies, it is anticipated that DNA will ultimately be collected from a total of 1.6 million participants (1.2 million samples have already been collected). The differences between these 14 studies and the remaining 39 were explored. The 14 studies differed from the rest for all characteristics outlined in Table 2 with the exception of the 'targeted number of participants' characteristic. For example, they all targeted individuals, all except one began after the year 2000, and they used a computer more frequently as the administration mode for questionnaires.

In a second illustrative example, 'Occurrence of diabetes (participant)', 'Occurrence of diabetes in the family' and 'Body mass index' could be successfully recreated by 11 studies. If gene–environment interactions were to be examined, biological samples are expected to be collected from 848 000 participants in these 11 studies.

Discussion

The current article provides an overview of the potential for the DataSHaPER approach to serve as foundation for retrospective harmonization across large population-based studies. There can be little doubt that the overall harmonization observed seems good enough to provide real scientific utility. However, the potential for harmonization is influenced by the types of variables targeted and study characteristics.

Table 2 Univariate and multivariate models; study characteristics and pairing results

		Average across studies of the percentage of variables presenting as:				Generalized estimating equations analysis[a]	
	Number of studies	Complete match	Partial Proximate match	Partial Tentative match	Impossible match	β	SE
Univariate analysis							
Study design							
Cohort (0)	48	36	3	14	48		
Cross-sectional (1)	5	40	2	15	44	0.124	0.1384
Scientific focus							
Broad scientific focus (0)	42	36	2	15	46		
Focus on specific outcome or exposure (1)	11	35	3	9	53	−0.048	0.243
Population sample source							
General population (0)	44	38	3	14	46		
Specific subpopulation (1)	9	29	3	12	57	−0.369	0.2343
Participation history							
New participant (0)	46	37	3	14	47		
From pre-existing study (1)	7	32	2	13	53	−0.230	0.1447
Observational units							
Individuals (0)	47	37	3	14	46		
Families (1)	6	29	2	12	57	−0.358**	0.1469
Sex							
One sex only (0)	7	29	5	13	54		
Both sexes (1)	46	37	2	14	47	0.268	0.2895
Region of residence of the participants							
Europe (1)	26	35	2	15	48		
North America (2)	17	40	4	13	44	0.298	0.225
Other (3)	10	33	1	13	53	−0.122	0.1572
Lower limit of age at recruitment (years)							
25–40 (1)	19	38	3	15	45		
<25 (2)	25	36	2	13	49	−0.084	0.2075
>40 (3)	9	34	3	16	48	−0.172	0.2838
Data collection start date							
Before or in 2000 (0)	23	27	3	14	56		
After 2000 (1)	30	43	2	14	41	0.639***	0.1464
Targeted number of participants							
Less than 100000 (0)	38	36	3	14	48		
More than or equal to 100000 (1)	15	37	3	14	46	0.086	0.1978
Proportion of biological samples collected							
Subsample (0)	23	29	3	14	54		
All participants (1)	30	41	2	14	43	0.461**	0.1589
Mode of administration of the questionnaire							
Computer-based (1)	13	46	2	13	39		
Paper-based (2)	24	29	2	15	54	−0.711***	0.1902
Mix (3)	16	38	3	14	45	−0.271	0.2119

(continued)

Table 2 Continued

	Number of studies	Average across studies of the percentage of variables presenting as: Complete match	Partial Proximate match	Partial Tentative match	Impossible match	Generalized estimating equations analysis[a] β	SE
Site of administration of the questionnaires							
Participant residence (1)	20	31	3	15	51		
Clinic/assessment centre (2)	12	38	2	16	45	0.23	0.1742
Mix (3)	21	39	3	13	45	0.331	0.2087
Multivariate analysis							
Observational units							
Individuals (0)						–	–
Families (1)						−0.401***	0.1047
Data collection start date							
Before or in 2000 (0)						–	–
After 2000 (1)						0.460**	0.1350
Mode of administration							
Computer-based (1)						–	–
Paper-based (2)						−0.548**	0.1797
Mix (3)						−0.275	0.2056

[a]Structure of the analysis: complete or partial proximate vs partial tentative or impossible.
Figures quoted denote log ORs, robust standard errors and associated *P*-values (*significant at <0.05, **significant at <0.01, ***significant at <0.001).

Recognition of an 'acceptable but restricted' level of heterogeneity is inherent to the DataSHaPER approach. Harmonization potential would be much lower if the use of 'identical' data collection procedures and tools were required to synthesize data. As an illustrative example, 16 (30%) of the participating studies used the specific wording 'Has a doctor (or physician) ever told you that you had diabetes?' in their questionnaire. The flexibility provided by the DataSHaPER approach increased the potential for harmonization to 89% (47 studies allowed construction of the variable 'Occurrence of diabetes'). As an example of flexibility, even if two distinct questions for types 1 and 2 diabetes were used, the variable 'Occurrence of diabetes' could be constructed. However, it can be hazardous to achieve harmonization procedures without using 'systematic-' and 'scientifically-based' rules and procedures. For example, the potential to harmonize the question 'Has a health professional ever told you that you have diabetes or high blood sugar levels?' could be interpreted differently by two independent investigators. In order to ensure quality of the harmonization process and, thus, the quality and scientific usefulness of the final database generated, it is essential to follow rigorous procedures and to have access to systematic and comprehensive information on the data collection procedures adopted by the studies involved.

Although this project has helped demonstrate the value of the DataSHaPER, some of its limitations must be highlighted. These limitations can be related to the: (i) variables selected and pairing rules defined; (ii) participating studies; and (iii) harmonization process. First, even if defined under consensus workshops, there is inevitably an element of subjectivity in variable selection and pairing rules definition. Furthermore, the Generic DataSchema was developed in order to support the harmonization of general purpose baseline questionnaires and physical measures. The variables and pairing rules have been selected and defined in that context. Therefore, they will not be appropriate for 'all' scientific questions that could potentially be addressed using the harmonized information. DataSchemas developed by new research programmes could certainly make use of the variables and rules already developed, but investigators will have to ensure proper customization of the variables and rules, or develop completely new DataSchemas to reflect their specific needs. The second point refers to participating studies. A relatively good participation

Table 3 Univariate and multivariate models; variable characteristics and pairing results

		Average across studies of the percentage of variables presenting as:				Generalized estimating equations analysis[a]	
	Number of variables	Complete match	Partial Proximate match	Partial Tentative match	Impossible match	β	SE
Univariate analysis							
Variable importance							
Essential (1)	38	60	2	15	23		
Important (2)	45	38	5	13	44	−0.776***	0.0932
Useful (3)	65	21	1	14	64	−1.772***	0.1239
Targeted individual							
Participant (0)	101	46	3	12	39		
Participant's family members (1)	47	14	1	19	66	−1.679***	0.199
Targeted period							
Current status (0)	40	50	4	11	36		
All other periods (1)	108	31	2	15	52	−0.841***	0.1112
Multivariate analysis							
Variable importance							
Essential (1)						–	–
Important (2)						−0.515***	0.0859
Useful (3)						−1.214***	0.0987
Targeted individual							
Participant (0)						–	–
Participant's family members (1)						−1.114***	0.1883
Targeted period							
Currently status (0)						–	–
All other periods(1)						−0.245*	0.1163

[a]Structure of the analysis: complete or partial proximate versus partial tentative or impossible.
Figures quoted denote log odds ratios, robust standard errors and associated *P*-values (*significant at <0.05, **significant at <0.01, ***significant at <0.001).

rate was observed. However, sampling was ultimately based on a comprehensive, but finite number of studies and specific selection criteria (e.g. availability of questionnaires in French or English). Furthermore, a higher participation rate was observed for studies launched after 2000. Consequently, participating studies should not be considered as representative of 'all' studies with similar profiles. The third point relates to the quality of the harmonization process. A rigorous approach has been used in order to achieve harmonization. However, as the current exercise was the first harmonization project to make use of the DataSHaPER tools, procedures and web interfaces were necessarily tailored as the project proceeded. Furthermore, even though the present article formally evaluates the harmonization potential of the DataSHaPER, actual pooling of data was not achieved and therefore quality of a common database cannot be assessed. Further work is therefore underway in order to formally validate the quality of real data that have been synthesized using the DataSHaPER.

In addition to the integration of formal procedures to assess the quality of synthesized databases generated, several additional developments will be required in the near future, if we are to optimize the utility of the DataSHaPER. For example, the current article focuses on information collected from participants at baseline interview, but it is essential to extend the harmonization process to longitudinal data. The DataSHaPER software and ontology is currently customized to integrate repeated assessments in order to respond to that issue. Furthermore, harmonization was initially limited to questionnaires and physical measures. To broaden the scope and to increase the

Table 4 Case study: number of studies and participants that could be included in a common analysis

Co-analysed variable(s)[a]	Number of studies for which variable provides a 'Complete' or 'Partial Proximate' match	Targeted number of participants	Number of current participants	Targeted number of participants with biological samples
(1) Systolic blood pressure	31 studies	3 731 000	1 816 000	3 326 000
(2) Diastolic blood pressure				
(1) Systolic blood pressure	28 studies	3 131 000	1 700 000	2 746 000
(2) Diastolic blood pressure				
(3) Body mass index				
(4) Level of physical activity				
(1) Systolic blood pressure	14 studies	2 019 000	1 335 000	1 649 000
(2) Diastolic blood pressure				
(3) Body mass index				
(4) Level of physical activity				
(5) Current use of alcohol				
(6) Current quantity of cigarettes smoked				
(7) Some elements of post-secondary education				

[a]Age and sex are also collected by all studies.

utility of the platform, the DataSHaPER is also being customized to integrate information extracted from additional sources, such as biospecimen repositories (e.g. samples processing and banking conditions) and registries (e.g. hospitalization, disease, death registries, environmental databases).

A number of more generic challenges must also be noted. First, important technical considerations need to be addressed to enable data synthesis. For example, the heterogeneity of the information management systems (e.g. Excel spreadsheet, SQL relational databases, SAS, SPSS, etc.) used by studies represents an important obstacle. To enable synthesis, a common data model that facilitates a consistent data representation across studies needs to be implemented in addition to the application of variable transformation rules (i.e. algorithms) that enable unit conversions and harmonized variable construction. Restrictive data-access policies and participant consent forms that prevent studies from sending individual data to third parties (e.g. a central data warehouse) also represent an important challenge for data pooling. However, there are ways to circumvent such problems, for example, by implementing a federated database network through methodological tools such as DataSHIELD[10] or using aggregated data in a study-level meta-analysis. Nevertheless, the potential for access to data and samples seems to be generally good for the participating studies. A survey has been conducted by P^3G to document the access policies of the 53 participating studies. To date, information has been provided by 45 of the studies and all but 4 permit access to data and samples by external researchers. However, whenever external access to information is permitted, limitations related to consent and intellectual property issues, the costing structure imposed and the time needed to navigate formal access procedures and receive data can all represent substantial burdens on scientists wishing to make use of information.

The generous collaborative involvement of some of the world's largest and most respected population-based studies has provided an ideal platform to launch the DataSHaPER project and will provide invaluable expertise for its future development. This first collaborative analysis will allow the improvement of the Generic DataSchema and DataSHaPER open source software. Updates, taking into account the results, of the harmonization process have been posted on the web for open access (www.datashaper.org). The collaborative analysis also raised the prospect of using the DataSHaPER far more extensively and for a wide range of different scientific purposes. We continually look to involve more population-based studies, investigators and tool developers in the DataSHaPER initiative and will welcome any suggestions or new projects making use of the platform from readers of this article. It is only through the realization of collaborative projects undertaking applied research using harmonized data that the DataSHaPER tool will be improved and that the exciting research potential provided by synthesized databases can ultimately be achieved.

Funding

Genome Canada and Genome Quebec (The Public Population Project in Genomics); Canadian Partnership Against Cancer (CPT); European FP6 (LSHG-CT-2006- 518418 to Promoting Harmonization of Epidemiological Biobanks in Europe); Medical Research Council Project Grant (G0601625; methods programme in genetic epidemiology at the University of Leicester that focuses on genetic statistics and largescale data harmonization and pooling); Wellcome Trust Supplementary Grant (086160/Z/08/A); National Institute for Health Research (Leicester Biomedical Research Unit in Cardiovascular Science); J.L. is a Canada Research Chair in Human Genome Epidemiology.

Acknowledgements

We would like to thank all of the studies and biobanking experts that provided advice and information on the development of the DataSHaPER, and are now part of the ongoing collaboration that is taking the DataSHaPER project forward. The funders had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

Supplementary Data

Supplementary Data are available at *IJE* online.

Conflict of interest: None declared.

References

[1] Burton PR, Hansell AL, Fortier I *et al*. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2009;**38:**263–73.

[2] Spencer CC, Su Z, Donnelly P *et al*. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009;**5:**e1000477.

[3] Hindorff LA, Sethupathy P, Junkins HA *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009;**106:**9362–67.

[4] Hunter DJ. Gene-environment interactions in human diseases. *Nat Rev Genet* 2005;**6:**287–98.

[5] Wong MY, Day NE, Luan JA *et al*. The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *Int J Epidemiol* 2003;**32:**51–57.

[6] Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. *Eur J Epidemiol* 2009;**24:**727–31.

[7] Fortier I, Burton PR, Robson PJ *et al*. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010;**39:** 1383–89.

[8] DataSHaPER. *The DataSHaPER (DataSchema and Harmonization Platform for Epidemiological Research) 2009*. http://www.datashaper.org (23 June 2011, date last accessed).

[9] OBiBa. *Open Source Software for Biobanks 2010*. http://www.obiba.org/ (23 June 2011, date last accessed).

[10] Wolfson M, Wallace SE, Masca N *et al*. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol* 2010;**39:**1372–82.

[11] Gostev M, Fernandez-Banet J, Rung J *et al*. SAIL – a software system for sample and phenotype availability across biobanks and cohorts. *Bioinformatics* 2011;**27:** 589–91.

[12] Inventory.nl and The Groningen Bioinformatics Center. *MOLGENIS Project, 2010*. http://www.molgenis.org/ (23 June 2011, date last accessed).

[13] BioSHaRE. *Biobank Standardisation and Harmonisation for Research Excellence in the European Union 2011*. http://www.bioshare.eu/ (23 June 2011, date last accessed).

[14] Borugian MJ, Robson P, Fortier I *et al*. The Canadian Partnership for Tomorrow Project: building a pan-Canadian research platform for disease prevention. *CMAJ* 2010;**182:**1197–1201.

[15] Dalkey N, Helmer O. An experimental application of the Delphi method to the use of experts. *Manage Sci* 1963;**9:** 458–67.

[16] Glaser EM. Using behavioral science strategies for defining the state-of-the-art. *J App Behav Sci* 1980;**16:** 79–92.

[17] World Health Organization. *International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Version for 2007*. http://apps.who.int/classifications/apps/icd/icd10online/ (23 June 2011, date last accessed).

[18] Craig CL *et al*. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc* 2003;**35:**1381–95.

[19] International Labour Organization. *International Standard Classification of Occupations (ISCO), 2010*. http://www.ilo.org/public/english/bureau/stat/isco/index.htm (23 June 2011, date last accessed).

[20] Stolk RP, Rosmalen JG, Postma DS *et al*. Universal risk factors for multifactorial diseases: LifeLines: a three-generation population-based study. *Eur J Epidemiol* 2008;**23:**67–74.

[21] Almqvist C, Adami HO, Franks PW *et al*. Lifegene–a large prospective population-based study of global relevance. *Eur J Epidemiol* 2011;**26:**67–77.

[22] Raina PS, Wolfson C, Kirkland SA *et al*. The Canadian longitudinal study on aging (CLSA). *Can J Aging* 2009; **28:**221–9.

[23] Public Population Project in Genomics (P3G). *P3G Observatory - Questionnaire Keywords, 2010*. http://www.p3gobservatory.org/Observatory.html#QUESTIONNAIRE_KEYWORDS.

[24] Armitage P, Berry G. *Statistical Methods in Medical Research*. 3rd edn. Oxford: Blackwell Scientific Publications, 1994.

[25] Garner JB. The standard error of Cohen's Kappa. *Stat Med* 1991;**10:**767–75.

[26] Bonnardel P. *The Kappa.exe Program, 2001*. http://kappa.chez-alice.fr/Kappa_cohen.htm (23 June 2011, date last accessed).

[27] Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986;**73:**13–22.

[28] Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med* 1998;**17:** 1261–91.

Appendix

International Harmonization Initiative

Participating Biobanks/Studies (arranged alphabetically by biobank/study, and then alphabetically by last name) are given below.

45 and Up Study (The): Emily Banks,[1,2] Louisa Jorm[1,3]; Agricultural Health Study: Laura Beane-Freeman,[4] Jane A Hoppin[5]; Airwave Health Monitoring Study: Paul Elliott,[6] Deepa Singh[6]; Australasian Colorectal Cancer Family Study: John Hopper,[7] Australian Breast Cancer Family Study: John Hopper[7]; Black Women's Health Study: Lynn Rosenberg,[8] Julie R Palmer[8]; Canadian Health Measures Survey: Gary Catlin,[9] Michael Wolfson[9]; Canadian Longitudinal Study on Aging: Susan Kirkland,[10] Parminder Raina,[11] Christina Wolfson[12,13]; Cancer Prevention Study – 3: Alpa V Patel,[14] Michael J Thun[14]; Cancer Prevention Study – II Nutrition Cohort: Susan M Gapstur,[14] Michael J Thun[14]; CARTaGENE: Claude Laberge,[15]; Cohort of Swedish Men: Niclas Hakansson[16] Alicja Wolk[16]; Constance Project: Marcel Goldberg,[17] Marie Zins[17]; Canadian Partnership for Tomorrow Project: Marilyn Borugian,[18,19] Richard P Gallagher,[18,19] John McLaughlin,[20] Louise Parker,[21] John D Potter,[22] Paula Robson[23]; The General Suburban Population Study (GESUS): Christina Ellervik[24]; European Prospective Investigation into Cancer and Nutrition: Aurelio Barricarte,[25,26] Franco Berrino,[27] Heiner Boeing,[28] Marie-Christine Boutron-Ruault,[29, 30] H Bas Bueno-de-Mesquita,[31,32] Françoise Clavel-Chapelon,[29,30] Miren Dorronsoro,[33] Carlos A Gonzalez,[34] Goran Hallmans,[35] Rudolf Kaaks,[36] Kay-Tee Kaw,[37] Tim J Key,[38] Eiliv Lund,[39] Jonas Manjer,[40] Traci Mouw,[41] Carmen Navarro,[42,43] Kim Overvad,[44] Domenico Palli,[45] Salvatore Panico,[46] Petra HM Peeters,[41, 47] Elio Riboli,[41] Laudina Rodriguez,[48] Isabelle Romieu,[49] Maria-José Sánchez,[43,50] Nadia Slimani,[49] Anne Tjønneland,[51] Antonia Trichopoulou,[52, 53] Rosario Tumino,[54] Paolo Vineis[41, 55]; Estonian Genome Centre: Helen Alavere,[56] Andres Metspalu[56]; FINRISK 2002: Markus Perola[57,58]; Gazel Cohort Study (The): Marcel Goldberg,[17] Marie Zins[17]; Generation Scotland – Scottish Family Health Study: Pamela Linksted,[59] Andrew D Morris[60]; Genome Database of Latvian Population: Janis Klovins,[61] Linda Tarasova[61]; Japan Public Health Center-based Prospective Study – Cohort I: Manami Inoue,[62] Shoichiro Tsugane[62]; Japan Public Health Center-based Prospective Study – Cohort II: Manami Inoue,[62] Shoichiro Tsugane[62]; KORA-gen – Cooperative health research in the Region of Augsburg: Angela Döring,[63] H Erich Wichmann[64,65,66]; Kadoorie Study of Chronic Disease in China: Zhengming Chen,[67] Liming Li[68,69]; Kaiser Permanente Research Program on Genes, Environment, and Health: Catherine Schaefer,[70] Larry Walter[70]; Korean Multi-Centre Cohort I: Sue K Park,[71,72] Keun-Young Yoo[71]; LifeGene: Mikael Eriksson,[73] Nancy Pedersen[73]; LifeLines Cohort Study and Biobank: Joost Keers,[74] Bruce HR Wolffenbuttel[75]; Malmö Diet Cancer: Jonas Manjer,[40] Peter Nilsson[76]; Melbourne Collaborative Cohort Study: Dallas R English[77]; Mexican-American Cohort Study: Melissa Bondy,[78] Anna Wilkinson[78]; Moli-sani Project: Amalia De Curtis,[79] Licia Iacoviello[79]; Montreal Heart Institute Biobank: Marie-Pierre Dubé,[80] Nathalie Laplante,[80] Jean-Claude Tardif[80]; NIH-AARP Diet and Health Study: Arthur Schatzkin,[81] Yikyung Park[81]; NUgene Project: Rex L Chisholm,[82] Wendy A Wolf[82]; National Child Development Study (1958 British birth cohort): Jon Johnson[83]; National DNA Bank – BancoADN: Andrés Garcia-Montero,[84] Alberto Orfao[84]; National FINRISK Study 2007 (The): Markus Perola[57,58]; National Health and Nutrition Examination Survey 2001-2002: Jody E McLean,[85] Geraldine M McQuillan[85]; National Health and Nutrition Examination Survey III: Jody E McLean,[85] Geraldine M McQuillan[85]; Nord-Trøndelag Health Study (The): Kristian Hveem[86]; Norwegian Women and Cancer Study postgenome cohort: Vanessa Dumeaux,[39] Eiliv Lund[39]; Nutrinet-Santé: Pilar Galan,[87,88] Serge Hercberg[87]; Screening Across the Lifespan Twin Study: Patrik KE Magnusson[73]; Singapore Consortium of Cohort Studies: Kee-Seng Chia,[89] En Yun Loy[89]; Sister Study (The): Genetic and Environmental Risk Factors for Breast Cancer: Lisa A DeRoo,[90] Dale P Sandler[90]; Sweden Mamography Cohort: Niclas Hakansson,[16] Alicja Wolk[16]; Swedish Twin Study of Adults: Genes and Environments (The): Patrik KE Magnusson[73]; UK Biobank: a large-scale prospective epidemiological resource: Rory Collins,[91] Tim Peakman[92]; UK Women's Cohort Study: Victoria J Burley,[93] Janet Cade,[93] Darren C Greenwood[93]; Vitamins And Lifestyle Study: A cohort Study of Dietary Supplements and Cancer Risk; Carolyn M Hutter,[94] Emily White[94]; Western Australia Sleep Health Study: Lyle J Palmer[95,96]; DataSHaPER development team: François L'Heureux,[97] Geneviève Lachance,[97] Cédric Thiebault,[97] Anne Vilain,[97] Mayss Naccache,[97] Ferima Sanogo,[97] Étienne Morency-Bachand,[97] Clément Tamisier,[97] Susan A Atkinson[98] and Andrea Rengifo,[98] Mylène Deschênes[97]

[1]The Sax Institute, Sydney, Australia, [2]National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australia, [3]School of Medicine, University of Western Sydney, Campbelltown Campus, Penrith South, Australia, [4]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA, [5]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC, USA, [6]Department of Epidemiology and Biostatistics, School of Public Health, and MRC-HPA Centre for Environment and Health, Imperial College London, St Mary's Campus, London, UK, [7]Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne, Melbourne, Victoria, Australia, [8]Slone Epidemiology Center, Boston University, Boston, MA, USA, [9]Statistics Canada, Ottawa, Ontario, Canada, [10]Department of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada, [11]Department of Clinical Epidemiology and Biostatistics, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada, [12]Division of Clinical Epidemiology, Research Institute McGill University Health Centre Professor, Montreal, Quebec, Canada, [13]Department of Epidemiology & Biostatistics & Occupational Health, and Department of Medicine, McGill University, Montreal, Quebec, Canada, [14]Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, GA, USA, [15]Faculty of Medicine, Laval University, Québec, Canada, [16]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden, [17]Inserm-Versaillles Saint Quentin University, Unité 1018, Hopital Paul Brousse, Villejuif, France, [18]Cancer Control Research, British Columbia Cancer Agency, Vancouver, BC, Canada, [19]Department of Health Care and Epidemiology, University of British Columbia, Vancouver, BC, Canada, [20]Samuel Lunenfeld Research Institute of the Mount Sinai Hospital, Toronto, ON, Canada, [21]Department of Medicine and Department of Paediatrics, Dalhousie University, Halifax, NS, Canada, [22]Canacer Prevention Program, Division of Public Health Sciences, Fred Hutchison Cancer Research Centre, Seattle, WA, USA, [23]Population Health Research, Alberta Health Services – Cancer Care, Edmonton, AB, Canada, [24]Department of Clinical Biochemistry, Copenhagen University Hospital, Naestved, Denmark, [25]Navarre Institute Public Health, Pamplona, Spain, [26]Consortium for Biomedical Research in Epidemiology and Public Health (CIBER Epidemiología y Salud Pública-CIBERESP), Spain, [27]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale Tumori, Milano, Italy, [28]Department of Epidemiology, German Institute of Human Nutrition, Potsdam-Rehbruecke, Germany, [29] Inserm, Centre for Research in Epidemiology and Population Health, U1018, Institut Gustave Roussy, F-94805, Villejuif, France, [30]Paris South University, UMRS 1018, F-94805, Villejuif, France, [31]National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands, [32]Department of Gastroenterology and Hepatology, University Medical Center Utrecht (UMCU), Utrecht, The Netherlands, [33]Public Health Department of Gipuzkoa, San Sebastian, Spain, [34]Unit of Nutrition, Environment and Cancer, Catalan Institute of Oncology, IDIBELL, Barcelona, Spain, [35]Department of Public Health and Clinical Medicine, Nutritional Research, Umeå University, Umeå, Sweden, [36]Division of Clinical Epidemiology, German Cancer Research Center, Heidelberg, Germany, [37]Clinical Gerontology, University of Cambridge, UK, [38]Cancer Epidemiology Unit, University of Oxford, Oxford, UK, [39]Department of Community Medicine, University of Tromsø, Tromsø, Norway, [40]Department of Surgery, Malmö University Hospital, Malmö, Sweden, [41]Department of Epidemiology & Public Health, Imperial College London, UK, [42]Epidemiology Department, Regional Health Authority, Murcia, Spain, [43]CIBER en Epidemiología y Salud Pública (CIBERESP), Spain, [44]Department of Cardiology and Department of Clinical Epidemiology, Aarhus University Hospital, Aalborg, Denmark, [45]Molecular and Nutritional Epidemiology Unit, CSPO-Scientific Institute of Tuscany, Florence, Italy, [46]Dipartimento di Medicina Clinica e Sperimentale, Università di Napoli, Italy, [47]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands, [48]Health Information Unit, Public Health and Health Planning Directorate, Asturias, Spain, [49]International Agency for Research on Cancer (IARC-WHO), Lyon, France, [50]Escuela Andaluza de Salud Publica, Granada, Spain, [51]Danish Cancer Society, Institute of Cancer Epidemiology, Copenhagen, Denmark, [52]Department of Hygiene and Epidemiology, School of Medicine, University of Athens, Greece, [53]Hellenic Health Foundation, Athens, Greece, [54]Cancer Registry, Azienda Ospedaliera ''Civile M.P.Arezzo'', Ragusa, Italy, [55]University of Torino, Torino, Italy, [56]Estonian Genome Project of University of Tartu, Tartu, Estonia, [57]National Institute for Welfare and Health, Helsinki, Finland, [58]Institute for Molecular Medicine Finland FIMM, University of Helsinki and National Public Health Institute, Helsinki, Finland, [59]Generation Scotland, University of Edinburgh, Molecular Medicine Centre, Western, General Hospital, Edinburgh, UK, [60]Biomedical Research Institute, University of Dundee, UK, [61]Latvian Biomedical Research and Study Center, Riga, Latvia, [62]Epidemiology and Prevention Division, Research Center for Cancer Prevention and Screening, National Cancer Center, Tokyo, Japan, [63]GSF National Research Center for Environment and Health, Institute of Epidemiology, Neuherberg, Germany, [64]Institute of Epidemiology, Helmholtz Zentrum München, Ludwig-Maximilians-Universität, Munich,

Germany, [65]Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany, [66]Klinikum Grosshadern, Ludwig-Maximilians-Universität München, Munich, Germany, [67]Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), University of Oxford, Oxford, UK, [68]The Chinese Academy of Medical Sciences, Beijing, China, [69]School of Public Health, Peking University, Beijing, China, [70]Research Program on Genes, Environment and Health, Kaiser Permanente Division of Research, Oakland, CA, USA, [71]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Korea, [72]Cancer Research Institute, Seoul National University Institute of Health Policy and Management, Seoul National University, Seoul, Korea, [73]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, [74]LifeLines, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands, [75]Department of Endocrinology and Metabolism, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands, [76]Department of Clinical Sciences, Malmö University Hospital, Lund University, Malmö, Sweden, [77]Center for Genetic Epidemiology, University of Melbourne, Melbourne, Australia, [78]Department of Epidemiology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA, [79]Laboratory of Genetic and Environmental Epidemiology, Research Laboratories, ''John Paul II'' Center for High Technology Research and Education in Biomedical Sciences, Catholic University, Campobasso, Italy, [80]Montreal Heart Institute, Université de Montréal, Montréal, Quebec, Canada, [81]Nutritional Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA, [82]The Center for Genetic Medicine, Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL, USA, [83]Centre for Longitudinal Studies, Institute of Education, University of London, London, UK, [84]Banco Nacional de ADN, Universidad de Salamanca, Fundacion Genoma España, Consejería de Sanidad de la Junta de Castilla y León, Spain, [85]Division of Health and Nutrition Examination Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA, [86]Department of Public Health, Norwegian University of Science and Technology, Trondheim, Norway, [87]UMR U 557 INSERM, U 1125 INRA, CNAM, Université Paris 13, Bobigny, France, [88]Département de Santé Publique, Hôpital Avicenne, Bobigny, France, [89]Department of Epidemiology and Public Health, Centre for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, [90]Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA, [91]Clinical Trial Service Unit and Epidemiological Studies Unit, University of Oxford, Oxford, UK, [92]UK Biobank, Units 1&2 Spectrum Way, Adswood, Stockport, Cheshire, UK, [93]Centre for Epidemiology and Biostatistics, Nutritional Epidemiology Group, University of Leeds, Leeds, UK, [94]Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [95]Genetic Epidemiology and Biostatistics Platform, Ontario Institute for Cancer Research, Toronto, Ontario, Canada, [96]Samuel Lunenfeld Research Institute, University of Toronto, Toronto, Canada, [97]Public Population Project in Genomics (P3G), Montreal, QC, Canada, [98]Department of Health Sciences and Department of Genetics, University of Leicester, Leicester, UK